Statistics 512: Applied Linear Models

Topic 5

Topic Overview

This topic will cover

- Diagnostics and Remedial Measures
- Influential Observations and Outliers

Chapter 10: Regression Diagnostics

We now have more complicated models. The ideas (especially with regard to the residuals) of Chapter 3 still apply, but we will also concern ourselves with the detection of outliers and influential data points. The following are often used for the identification of such points and can be easily obtained from SAS:

- Studentized deleted residuals
- Hat matrix diagonals
- Dffits, Cook's D, DFBETAS
- Variance inflation factor
- Tolerance

Life Insurance Example

- We will use this as a running example in this topic.
- References: page 386 in KNNL and `nknw364.sas`.
- Y = amount of insurance (in \$1000)
- X_1 = Average Annual Income (in \$1000)
- X_2 = Risk Aversion Score (0-10)
- $n = 18$ managers were surveyed.

```
data insurance;
infile 'H:\System\Desktop\Ch09ta01.dat';
input income risk amount;
proc reg data=insurance;
   model amount=income risk/r influence;
```

Just to get oriented...

Analysis of Variance

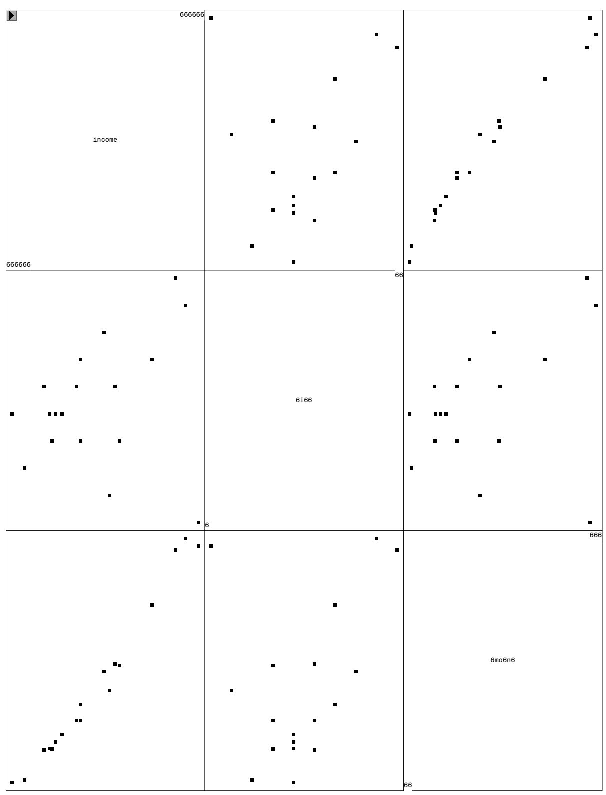
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	173919	86960	542.33	<.0001
Error	15	2405.14763	160.34318		
Corrected Total	17	176324			

Root MSE	12.66267	R-Square	0.9864
Dependent Mean	134.44444	Adj R-Sq	0.9845
Coeff Var	9.41851		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-205.71866	11.39268	-18.06	<.0001
income	1	6.28803	0.20415	30.80	<.0001
risk	1	4.73760	1.37808	3.44	0.0037

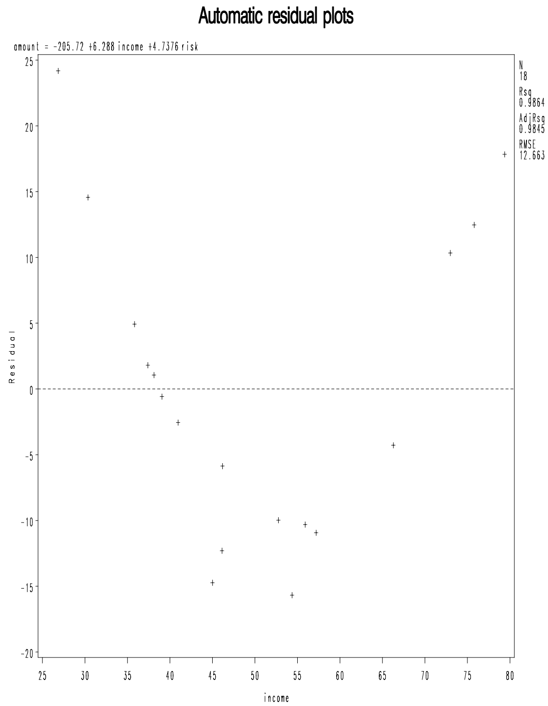
Model is significant and $R^2 = 0.9864$ – quite high – both variables are significant.

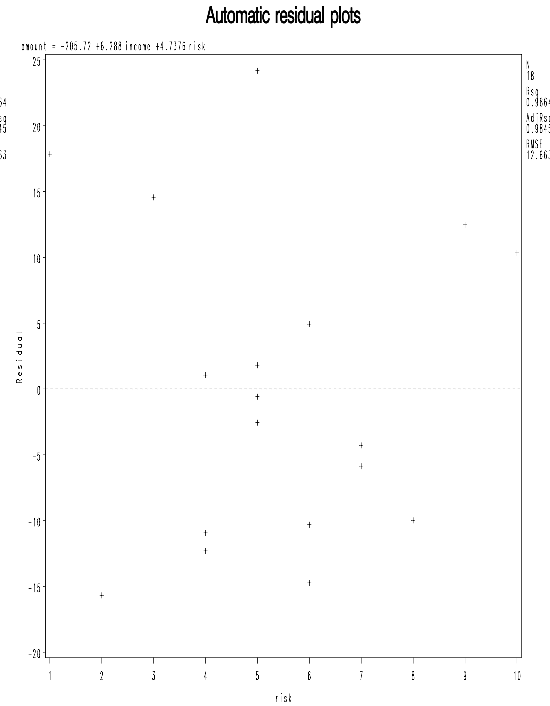


The Usual Residual Plots

The `plot` statement generates the following two residual plots (in the past we have used `gplot` to create these). These residuals are for the full model. Note the weird syntax r.*(income risk). It prints the estimated equation and the R^2 on it automatically, which is kind of nice. This is an alternative to saving the residuals and using gplot, although you have less control over the output.

```
title1 'Insurance';
proc reg data=insurance;
   model amount=income risk/r partial;
   plot r.*(income risk);
```





It looks like there is something quadratic going on with *income* in the full model. The residuals for *risk* look okay.
(We should also do a qqplot.)

Types of Residuals

Regular Residuals

- $e_i = Y_i - \hat{Y}_i$ (the usual).
- These are given in the SAS output under the heading "Residual" when you use the r option in the model statement, and to store them use r = (name) in an output statement.

Studentized Residuals

- $e_i^* = \frac{e_i}{\sqrt{MSE \times (1-h_{i,i})}}$

- *Studentized* means divided by its standard error. (When you ignore the $h_{i,i}$ and just divide by `Root MSE` they are called *semistudentized residuals*.)
- Recall that $s^2\{\mathbf{e}\} = MSE(\mathbf{I} - \mathbf{H})$, so that $s^2\{e_i\} = MSE(1 - h_{i,i})$. These follow a $t_{(n-p)}$ distribution if all assumptions are met.
- Studentized residuals are shown in the SAS output under the heading "`Student Residual`." In the output, "`Residual`" / "`Std Error Residual`" = "`Student Residual`". SAS also prints a little bar graph of the studentized residuals so you can identify large ones quickly.
- In general, values larger than about 3 should be investigated. (The actual cutoff depends on a t distribution and the sample size; see below.) These are computed using the '`r`' option and can be stored using `student=(name)`.

Studentized Deleted Residuals

- The idea: delete case i and refit the model. Compute the predicted value and residual for case i using this model. Compute the "studentized residual" for case i. (Don't do this literally.)
- We use the notation (i) to indicate that case i has been deleted from the computations.
- $d_i = Y_i - \hat{Y}_{i(i)}$ is the deleted residual. (Also used for PRESS criterion)
- Interestingly, it can be calculated from the following formula without re-doing the regression with case i removed. It turns out that $d_i = \frac{e_i}{(1-h_{i,i})}$, where $h_{i,i}$ is the ith diagonal element of the Hat matrix $\mathbf{H}$. Its estimated variance is $s^2\{d_i\} = \frac{MSE_{(i)}}{(1-h_{i,i})}$.
- The studentized deleted residual is $t_i = \frac{d_i}{\sqrt{s^2\{d_i\}}} = \frac{e_i}{(1-h_{i,i})}\sqrt{\frac{(1-h_{i,i})}{MSE_{(i)}}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{i,i})}}$.
- $MSE_{(i)}$ can be computed by solving this equation: $(n-p)MSE = (n-p-1)MSE_{(i)} + \frac{e_i^2}{1-h_{i,i}}$.
- The t_i are shown in the SAS output under the heading "`Rstudent`", and the $h_{i,i}$ under the heading "`Hat Diag H`". To calculate these, use the `influence` option and to store them use `rstudent=(name)`.
- We can use these to test (using a Bonferroni correction for n tests) whether the case with the largest studentized residual is an outlier (see page 396).

```
proc reg data=insurance;
   model amount=income risk/r influence;
```

Output Statistics

Obs	Dep Var amount	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	91.0000	-14.7311	12.216	-1.206	| **| |	0.036
2	162.0000	-10.9321	12.009	-0.910	| *| |	0.031
3	11.0000	24.1845	11.403	2.121	| |**** |	0.349
4	240.0000	-4.2780	11.800	-0.363	| | |	0.007
5	73.0000	-2.5522	12.175	-0.210	| | |	0.001
6	311.0000	10.3417	10.210	1.013	| |** |	0.184
7	316.0000	17.8373	7.780	2.293	| |**** |	2.889
8	154.0000	-9.9763	11.798	-0.846	| *| |	0.036
9	164.0000	-10.3084	12.239	-0.842	| *| |	0.017
10	54.0000	1.0560	12.009	0.0879	| | |	0.000
11	53.0000	4.9301	11.878	0.415	| | |	0.008
12	326.0000	12.4728	10.599	1.177	| |** |	0.197
13	55.0000	1.8081	12.050	0.150	| | |	0.001
14	130.0000	-15.6744	11.258	-1.392	| **| |	0.171
15	112.0000	-5.8634	12.042	-0.487	| | |	0.008
16	91.0000	-12.2985	12.162	-1.011	| **| |	0.029
17	14.0000	14.5636	11.454	1.271	| |** |	0.120
18	63.0000	-0.5798	12.114	-0.0479	| | |	0.000

Test for Outliers Using Studentized Deleted Residuals

- should use the Bonferroni correction since you are looking at all n residuals
- studentized deleted residuals follow a $t_{(n-p-1)}$ distribution since they are based on $n-1$ observations
- If a studentized deleted residual is bigger in magnitude than $t_{n-p-1}(1-\frac{\alpha}{2n})$ then we identify the case as a possible outlier based on this test.
- In our example, take $\alpha = 0.05$. Since $n = 18$ and $p = 3$, we use $t_{14}(0.9986) \approx 3.6214$.
- None of the observations may be called an outlier based on this test.
- Note that if we neglected to use the Bonferroni correction our cutoff would be 2.1448 which would detect obs. 3 and 7, but this would not be correct.
- Note that "identifying an outlier" does not mean that you then automatically remove the observation. It just means you should take a closer look at that observation and check for reasons why it should possibly be removed. It could also mean that you have problems with normality and/or constant variance in your dataset and should consider a transformation.

What to Look For

When we examine the residuals we are looking for

- Outliers

- Non-normal error distributions
- Influential observations

Other Measures of Influential Observations

The `influence` option calculates a number of other quantities. We won't spend a whole lot of time on these, but you might be wondering what they are.

```
                                        Output Statistics
          Cook's Hat Diag                ----------DFBETAS----------
Obs            D         H     DFFITS  Intercept     income       risk
  1        0.036    0.0693    -0.3345    -0.1179     0.1245    -0.1107
  2        0.031    0.1006    -0.3027    -0.0395    -0.1470     0.1723
  3        0.349    0.1890     1.1821     0.9594    -0.9871     0.1436
  4        0.007    0.1316    -0.1369     0.0770    -0.0821    -0.0410
  5        0.001    0.0756    -0.0580    -0.0394     0.0286     0.0011
  6        0.184    0.3499     0.7437    -0.5298     0.3048     0.5125
  7        2.889    0.6225     3.5292    -0.3649     2.6598    -2.6751
  8        0.036    0.1319    -0.3263     0.0816     0.0254    -0.2452
  9        0.017    0.0658    -0.2212     0.0308    -0.0672    -0.0366
 10        0.000    0.1005     0.0284     0.0238    -0.0138    -0.0092
 11        0.008    0.1201     0.1490     0.0863    -0.1057     0.0536
 12        0.197    0.2994     0.7801    -0.5820     0.4495     0.4096
 13        0.001    0.0944     0.0468     0.0348    -0.0294     0.0014
 14        0.171    0.2096    -0.7423    -0.2706    -0.2656     0.6269
 15        0.008    0.0957    -0.1543    -0.0164     0.0532    -0.0953
 16        0.029    0.0775    -0.2934    -0.1810     0.0258     0.1424
 17        0.120    0.1818     0.6129     0.5803    -0.3608    -0.2577
 18        0.000    0.0849    -0.0141    -0.0101     0.0080    -0.0001
 *         0.826    0.3333     0.8165     1 (or 0.4714)
```

Cook's Distance

- This measures the influence of case i on all of the $\hat{Y}_i$'s. It is a standardized version of the sum of squares of the differences between the predicted values computed with and without case i.
- Large values suggest an observation has a lot of influence. Cook's D values are obtained via the '`r`' option in the `model` statement and can be stored with `cookd=(name)`.
- here "large" means larger than the 50th percentile of the $F_{p,n-p}$ distribution; for our example $F_{3,15}(0.5) = 0.826$.

Hat Matrix Diagonals

- $h_{i,i}$ is a measure of how much Y_i is contributing to the prediction of $\hat{Y}_i$. This depends on the distance between the X values for the ith case and the means of the X values. Observations with extreme values for the predictors will have more influence.

- $h_{i,i}$ is sometimes called the *leverage* of the ith observation. It always holds that $0 \leq h_{i,i} \leq 1$ and $\sum h_{i,i} = p$.

- A large value of $h_{i,i}$ suggests that the ith case is distant from the center of all X's. The average value is p/n. Values far from this average (say, twice as large) point to cases that should be examined carefully because they may have a substantial influence on the regression parameters.

- For our example, $\frac{2p}{n} = \frac{6}{18} = 0.333$ so values larger than 0.333 would be considered large. Observations #6, #7, and maybe #12 seem to have a lot of influence. These can be further examined with the next set of influence statistics.

- The hat matrix diagonals are displayed with the `influence` option and can be stored with `h=(name)` .

`DEFITS`

- Another measure of the influence of case i on its own fitted value $\hat{Y}_i$. It is a standardized version of the difference between $\hat{Y}_i$ computed with and without case i. It is closely related to $h_{i,i}$ (consult the text for formula if you are interested). Values larger than 1 (for small to medium size datasets) or $2\sqrt{\frac{p}{n}}$ (for large datasets) are considered influential. (In our example, $2\sqrt{\frac{p}{n}} = 0.816$ but this is a small dataset so we would use 1).

- these are calculated with the `influence` option and can be stored with `dffits=(name)`.

`DFBETAS`

- A measure of the influence of case i on each of the regression coefficients.

- It is a standardized version of the difference between the regression coefficient computed with and without case i.

- Values larger than 1 (for small-to-medium datasets) or $\frac{2}{\sqrt{n}}$ (for large datasets) are considered influential. In this example $\frac{2}{\sqrt{n}} = 0.4714$, but we would use 1 as a cutoff.

- According to all these measures, observation #7 appears to be influential. This is not surprising because it has the smallest risk (1) and the highest income (79.380) of all the observations.

Measures of Multicollinearity

We already know about several identifying factors in dealing with multicollinearity:

- regression coefficients change greatly when predictors are included/excluded from the model

- significant F-test but *no* significant t-tests for β's (ignoring intercept)

- regression coefficients that don't "make sense", i.e. don't match scatterplot and/or intuition
- Type I and II SS very different
- predictors that have pairwise correlations

There are two other numerical measures that can be used: `vif` and `tolerance`

Variance Inflation Factor

- The VIF is related to the variance of the estimated regression coefficients.
- $VIF_k = \frac{1}{1-R_k^2}$ where R_k^2 is the squared multiple correlation obtained in a regression where all other explanatory variables are used to predict X_k. We calculate it for each explanatory variable.
- If this R_k^2 is large that means X_k is well predicted by the other X's. One suggested rule is that a value of 10 or more for VIF indicates excessive multicollinearity. This corresponds to an R_k^2 of ≥ 0.9. Use the `vif` option to the `model` statement.

Tolerance

- $TOL = 1 - R_k^2 = \frac{1}{VIF}$. A tolerance of < 0.1 is the same as a VIF > 10, indicating excessive multicollinearity. Use the `TOL` option to the `model` statement. Described in comment on p 388.

Typically you would look at either vif or tol, not both.

```
proc reg data=insurance;
   model amount=income risk/tol vif;
```

```
                                  Parameter Estimates
Variable     Tolerance     Inflation
Intercept            .             0
income         0.93524       1.06925
risk           0.93524       1.06925
```

These values are quite acceptable.

Partial Regression Plots

- Also called partial residual plots, added variable plots or adjusted variable plots.
- Related to partial correlations, they help you figure out the net effect of X_i on Y, given that other variables are in the model.

- One plot for each X_i. To get the plot, run two regressions. In the first, use the other X's to predict Y. In the second use the other X's to predict X_i. Then plot the residuals from the first regression against the residuals from the second regression. The correlation of these residuals was called the *partial correlation coefficient*.
- A linear pattern in this type of plot indicates that the variable would be useful in the model, and the slope is its regression coefficient. The plots shows the strength of a marginal relationship between Y and X_i in the full model. If the partial residual plot for X_i appears "flat", X_i may not need to be included in the model. If they appear like a straight line (with non-zero slope), then that suggests X_i should be included as a linear term, etc.
- Nonlinear relationships, heterogeneous variances, and outliers may also be detected in these plots.
- In SAS, the '`partial`' option in the `model` statement can be used to get a partial residual plot. This is not a very good plot (useful for first glance, but not something you would want to publish), so it is useful to know how to create a better one.

Coding for the poor resolution plot (they're kind of ugly):

```
proc reg data=insurance;
   model amount=income risk/r partial;
```

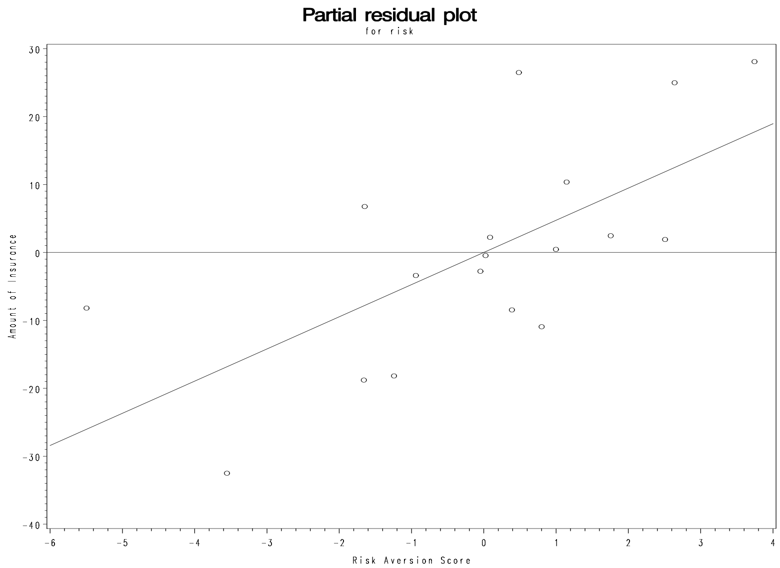
(The number labels on the plot are the first digit of income because we said "`id income`".) The axes are labelled `amount` and `income`, but we are actually plotting the residuals for *amount* (predicted by *risk*) vs. the residuals for *income* (when predicted by *risk*)

(The number labels on the plot are the first digit of *income* because we said "`id income`".)

Obtaining Partial Regression Plots

```
title1 'Partial residual plot';
title2 'for risk';
symbol1 v=circle i=rl;
axis1 label=('Risk Aversion Score');
axis2 label=(angle=90 'Amount of Insurance');
proc reg data=insurance;
   model amount risk = income;
   output out=partialrisk r=resamt resrisk;
proc gplot data=partialrisk;
   plot resamt*resrisk / haxis=axis1 vaxis=axis2 vref = 0;
run;
```

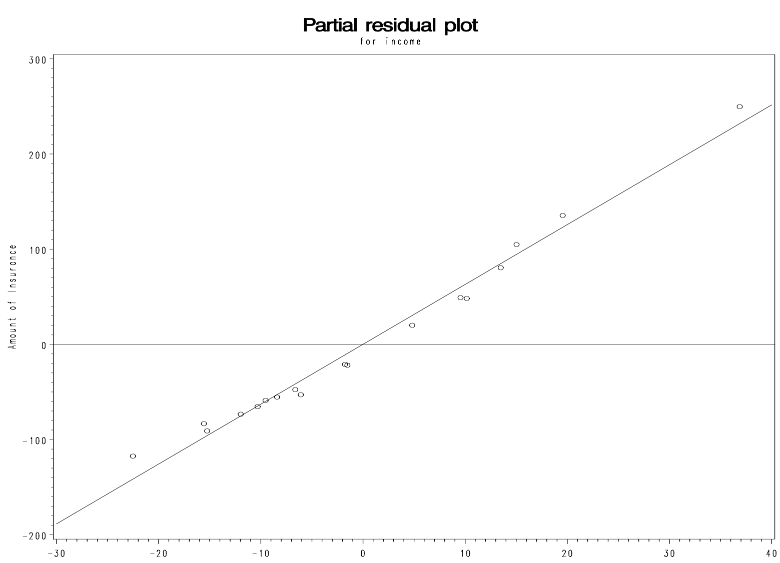
The y-axis has the residuals for the model `insur = income`. The x-axis has the residuals for the model `risk = income` (i.e. treat risk as a Y-variable).
The residuals compared to the horizontal line are the residuals for the model that omits *risk* as a variable. The residuals compared to the "regression" line are the residuals for the



model that includes *risk* as a variable. Are the points closer to the regression line than to the x-axis? This helps decide if there is much to be gained (i.e. smaller residuals) by including *risk* in the model. In this case *risk* clearly should be included.

Similar code for *income*:

```
axis3 label=('Income');
title2 'for income';
proc reg data=insurance;
model amount income = risk;
    output out=partialincome r=resamt resinc;
proc gplot data=partialincome;
   plot resamt*resinc / haxis=axis3 vaxis=axis2 vref = 0;
```



The resulting plot has on the y-axis the residuals for the model `insur = risk`, and the x-axis has the residuals for the model `income = risk`. This is the same as the text plot.

This plot shows, first of all, that *income* is clearly needed in the model. Secondly, we can see that the effect of *income* (when *risk* is included) is *mostly* linear. Third, a close look shows that the residuals curve a bit around the straight line, so that there is a quadratic effect. However, the quadratic effect is small compared to the linear one. A quadratic term will improve the fit of the model, but it may not improve it *much*. We would have to weigh the improved fit vs. the interpretability and possible multicollinearity problems when deciding on the final model.

Here's what happens when we include the square of (centered) income:

```
data quad;
    set insurance;
    sinc = income;
proc standard data=quad out=quad mean=0;
    var sinc;
data quad;
    set quad;
    incomesq = sinc*sinc;
title1 'Residuals for quadratic model';
proc reg data=quad;
    model amount = income risk incomesq / r vif;
    plot r.*(income risk incomesq);
```

Analysis of Variance

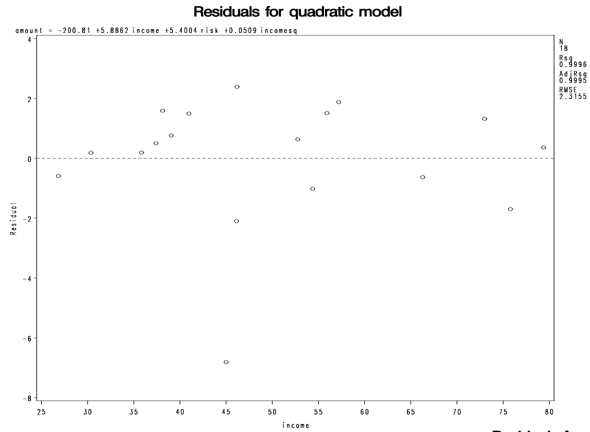
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	176249	58750	10958.0	<.0001
Error	14	75.05895	5.36135		
Corrected Total	17	176324			

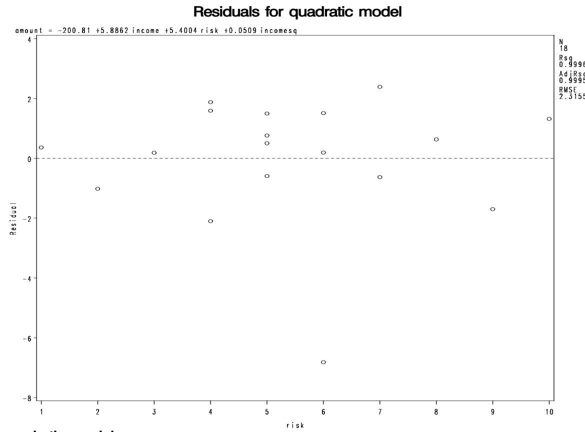
Root MSE	2.31546	R-Square	0.9996
Dependent Mean	134.44444	Adj R-Sq	0.9995
Coeff Var	1.72224		

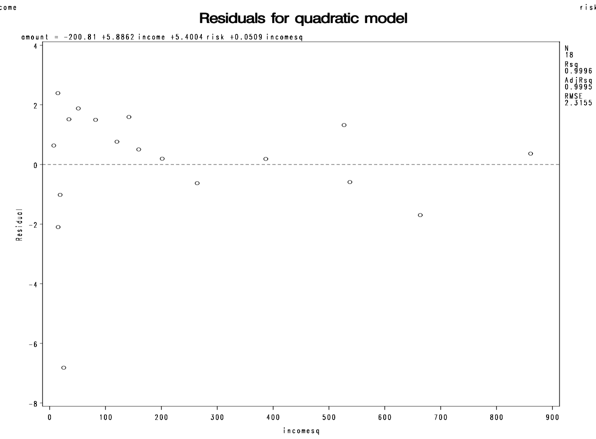
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-200.81134	2.09649	-95.78	<.0001	0
income	1	5.88625	0.04201	140.11	<.0001	1.35424
risk	1	5.40039	0.25399	21.26	<.0001	1.08627
incomesq	1	0.05087	0.00244	20.85	<.0001	1.26657

For the two-variable model, R^2 was 0.9864, so while this is an improvement, it does not make a big difference. Our assumptions are now more closely met, which is good, but it also appears an outlier now exists where it did not before.







Regression Diagnostics Summary

Check normality of the residuals with a normal quantile plot.
Plot the residuals versus predicted values, versus each of the X's and (when appropriate) versus time
Examine the partial regression plots for each X variable.
Examine

- the studentized deleted residuals (`RSTUDENT` in the output)
- The hat matrix diagonals
- Dffits, Cook's D, and the DFBETAS
- Check observations that are extreme on these measures relative to the other observations
- Examine the tolerance or VIF for each X

If there are variables with low tolerance / high VIF, or if any of the other indications of multicollinearity problems are present, you may need to do some model building:

- Recode variables
- Variable selection

Remedial Measures (Chapter 11)

- Weighted Regression
- Robust Regression
- Nonparametric Regression
- Bootstrapping

Weighted Regression

Maximum Likelihood

$$\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i, \ \text{Var}(\epsilon_i) = \sigma_i^2 \\
Y_i &\sim N(\beta_0 + \beta_1 X_i, \sigma_i^2) \\
f_i &= \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma_i}\right)^2} \\
L &= f_1 \times f_2 \times \cdots \times f_n \text{ – likelihood function}
\end{aligned}$$

- Variance is no longer constant
- Maximization of L with respect to β's.
- Equivalent to minimization of $\sum \frac{1}{\sigma_i^2}(Y_i - \beta_0 - \beta_1 X_{i,1} - \ldots - \beta_{p-1} X_{i,p-1})^2$

Weighted Least Squares

- Used to deal with unequal variances:

$$\sigma^{\mathbf{2}}\{\epsilon\} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

- Least squares minimizes the sum of the squared residuals. For WLS, we minimize instead the sum of the squared residuals each multiplied by an appropriate weight. If the error variances are known, the weights are $w_i = 1/\sigma_i^2$.
- Otherwise the variances need to be estimated (see discussion pages 403-405).
- The regression coefficients with weights are: $\mathbf{b_W} = (\mathbf{X'WX})^{-1}(\mathbf{X'WY})$ where $\mathbf{W}$ is a diagonal matrix of weights.
- In SAS, use a '`weight`' statement in `PROC REG`.

Drawbacks to Weighted Least Squares

No clear interpretation for MSE. MSE will be close to 1 if error variance is modeled well.

Advantages to Weighted Least Squares

Improved parameter estimates, and CI's. Valid inference in presence of heteroscedasticity.

Determining the Weights

We try to find a relationship between the absolute residual and another variable and use this as a model for the standard deviation; or similarly for the squared residual and the variance. Sometimes it is necessary to use grouped data or approximately grouped data to estimate the variance. With a model for the standard deviation or the variance, we can approximate the optimal weights. Optimal weights are proportional to the inverse of the variance as shown above. If the data have many observations for each value of X we can get a variance estimate at each value (this happens frequently in ANOVA).

KNNL Example

- KNNL p 427 (`nknw406.sas`)
- Y is diastolic blood pressure
- X is age
- $n = 54$ healthy adult women aged 20 to 60 years old

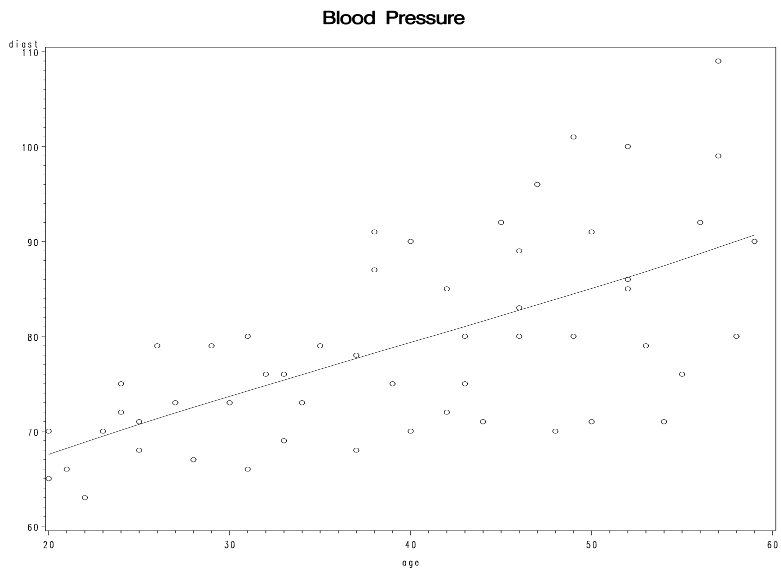
```
data pressure;
  infile 'H:\System\Desktop\Ch10ta01.dat';
  input age diast;
proc print data=pressure;
title1 'Blood Pressure';
symbol1 v=circle i=sm70;
proc sort data=pressure;
  by age;
proc gplot data=pressure;
  plot diast*age;
```

This clearly has non-constant variance. Run the (unweighted) regression to get residuals.

```
proc reg data=pressure;
  model diast=age / clb;
  output out=diag r=resid;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2374.96833	2374.96833	35.79	<.0001



Error	52	3450.36501	66.35317
Corrected Total	53	5825.33333	

Root MSE	8.14575	R-Square	0.4077
Dependent Mean	79.11111	Adj R-Sq	0.3963
Coeff Var	10.29659		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	56.15693	3.99367	14.06	<.0001	48.14304	64.17082
age	1	0.58003	0.09695	5.98	<.0001	0.38548	0.77458

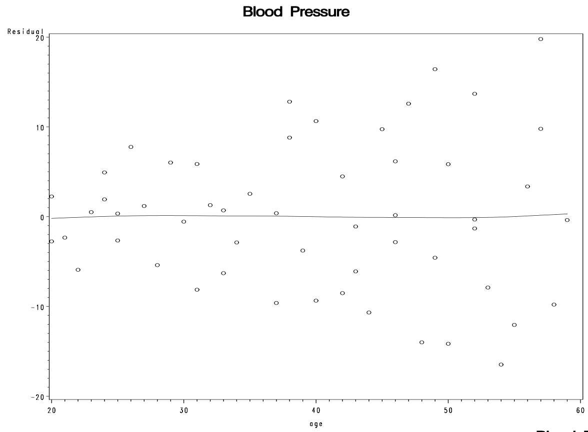
Use the output data set to get the absolute and squared residuals. Plot each of them (vs. X) with a smoother.

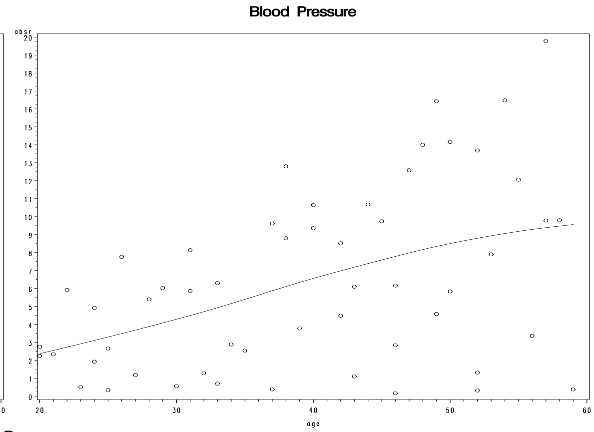
```
data diag;
  set diag;
  absr=abs(resid);
  sqrr=resid*resid;

proc gplot data=diag;
   plot (resid absr sqrr)*age;
```

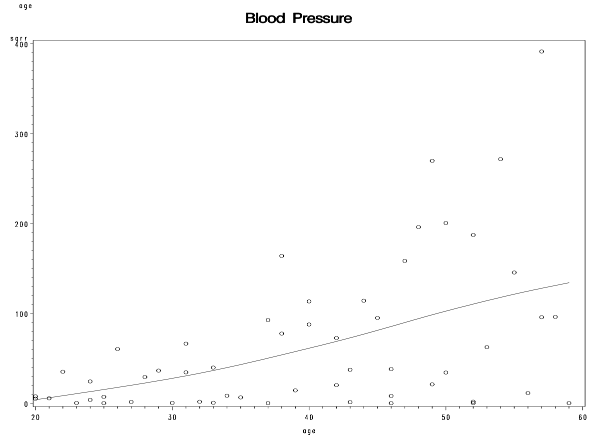
The absolute value of the residuals appears to have a fairly linear relationship with *age* (it appears more linear than does the graph of squared residuals vs. *age*). Thus, we will model standard deviation as a linear function of *age*. (If the second graph was more linear we would model variance instead.) We will model the absolute residuals as a function of *age*, and use the predicted values of that regression as weights.

Predict the standard deviation (absolute value of the residual):

```
proc reg data=diag;
```








```
   model absr=age;
   output out=findweights p=shat;
data findweights;
   set findweights;
   wt=1/(shat*shat);
```

We always compute the weights as the reciprocal of the estimated variance. Regression with weights:

```
proc reg data=findweights;
   model diast=age / clb p;
   weight wt;
   output out = weighted p = predict;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	83.34082	83.34082	56.64	<.0001
Error	52	76.51351	1.47141		
Corrected Total	53	159.85432			

Root MSE	1.21302	R-Square	0.5214
Dependent Mean	73.55134	Adj R-Sq	0.5122
Coeff Var	1.64921		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	55.56577	2.52092	22.04	<.0001	50.50718	60.62436
age	1	0.59634	0.07924	7.53	<.0001	0.43734	0.75534

Other Methods

Robust Regression

- Basic idea is to have a procedure that is not sensitive to outliers.
- Alternatives to least squares, minimize either the sum of absolute values of residuals or the median of the squares of residuals.
- Do weighted regression with weights based on residuals, and iterate.
- See Section 11.3 for details.

Nonparametric Regression

- Several versions
- We have used e.g. `i=sm70`
- Interesting theory
- All versions have some smoothing parameter similar to the `70` in `i=sm70`.
- Confidence intervals and significance tests not fully developed.

Bootstrap

- Very important theoretical development that has had a major impact on applied statistics
- Based on simulation
- Sample *with* replacement from the data or residuals and get the distribution of the quantity of interest
- CI usually based on quantiles of the sampling distribution

Model Validation

Three approaches to checking the validity of the model.

- Collect new data: does it fit the model?
- Compare with theory, other data, simulation.
- Use some of the data for the basic analysis ("training set") and some for validity check.

Qualitative Explanatory Variables (Section 8.3)

Example include

- Gender as an explanatory variable
- Placebo versus treatment
- Insurance Co. example from previous notes (Type of company)

Two Categories

Recall from Topic 4 (General Linear Tests):

- Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$
- When $X_1 = 0$, β_1 and β_3 terms disappear: $Y = \beta_0 + \beta_2 X_2 + \epsilon$. For this group, β_0 is the intercept, and β_2 is the slope.
- When $X_1 = 1$, β_1 and β_3 terms are incorporated into the intercept and X_2 coefficient:

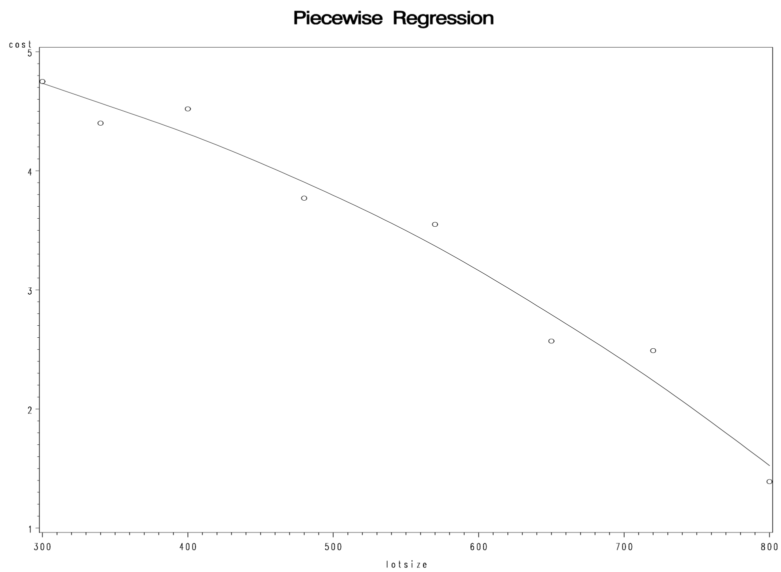
$$Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2 + \epsilon$$

- For this group, $\beta_0 + \beta_1$ is the intercept, and $\beta_2 + \beta_3$ is the slope.
- $\text{H}_0 : \beta_1 = \beta_3 = 0$ is the hypothesis that the regression lines are the same.
- $\text{H}_0 : \beta_1 = 0$ hypothesizes the two intercepts are equal.
- $\text{H}_0 : \beta_3 = 0$ hypothesizes the two slopes are equal.

More Complicated Models

- If a categorical (qualitative) variable has k possible values we need $k - 1$ indicator variables in order to describe it.
- These can be defined in many different ways; we will do this in Chapter 16 (ANOVA).
- We also can have several categorical explanatory variables, plus interactions, etc.
- Example: Suppose we have a variable *speed* for which 3 levels (high, medium, low) are possible. Then we would need two indicator variables (e.g. X_1 = medium and X_2 = high) to describe the situation.

speed	X_1	X_2
low	0	0
medium	1	0
high	0	1



Piecewise Linear Model

At some (known) point or points, the slope of the relationship changes. We can describe such a model with indicator variables.

Examples:

- tax brackets
- discount prices for bulk quantities
- overtime wages

Piecewise Linear Model Example

- Y = unit cost, X_1 = lot size, $n = 8$
- We have reason to believe that a linear model is appropriate, but a slope change should be allowed at $X_1 = 500$. (Note the 'bending' in the plot.)
- We can do this by including an indicator variable X_2 that is 1 if X_1 is bigger than 500 and 0 otherwise and allowing it to interact with X_1.

```
data piecewise;
   infile 'H:\System\Desktop\Ch11ta06.dat';
   input cost lotsize;
symbol1 v=circle i=sm70 c=black;
proc sort data=piecewise; by lotsize;
proc gplot data=piecewise;
   plot cost*lotsize;
```

Piecewise Model

Define a new variable X_2 which is 0 when $X_1 \leq 500$ and 1 when $X_1 > 500$. Then create an adjusted interaction term $X_3 = X_2(X_1 - 500)$. This uses $-500X_2$ to indicate the change in intercept and the product X_1X_2 to find the change in slope. Note that there is only one parameter since the two lines must join at $X_1 = 500$. We will not use X_2 explicitly in the model, just the interaction term X_3. Thus the model is

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \epsilon \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2(X_1 - 500) + \epsilon \\ &= \beta_0 - 500\beta_2 X_2 + \beta_1 X_1 + \beta_2 X_1 X_2 + \epsilon \\ &= \begin{cases} \beta_0 + \beta_1 X_1 & X_2 = 0 \quad (X_1 \leq 500) \\ (\beta_0 - 500\beta_2) + (\beta_1 + \beta_2) X_1 & X_2 = 1 \quad (X_1 > 500) \end{cases} \end{aligned}$$

Our model has

- An intercept (β_0)
- A coefficient for lot size (the slope β_1)
- An additional explanatory variable that will add a constant to the slope whenever lot size is greater than 500.

```
data piecewise; set piecewise;
   if lotsize le 500
     then cslope=0;
   if lotsize gt 500
     then cslope=lotsize-500;
proc print data=piecewise;
```

Obs	cost	lotsize	cslope
1	4.75	300	0
2	4.40	340	0
3	4.52	400	0
4	3.77	480	0
5	3.55	570	70
6	2.57	650	150
7	2.49	720	220
8	1.39	800	300

The variable `cslope` is our X_3. Run the regression:

```
proc reg data=piecewise;
   model cost=lotsize cslope;
   output out=pieceout p=costhat;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	9.48623	4.74311	79.06	0.0002

Error	5	0.29997	0.05999
Corrected Total	7	9.78620	

Root MSE	0.24494	R-Square	0.9693
Dependent Mean	3.43000	Adj R-Sq	0.9571
Coeff Var	7.14106		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.89545	0.60421	9.76	0.0002
lotsize	1	-0.00395	0.00149	-2.65	0.0454
cslope	1	-0.00389	0.00231	-1.69	0.1528

Plot data with fitted values:

```
symbol1 v=circle i=none c=black;
symbol2 v=none i=join c=black;
proc sort data=pieceout; by lotsize;
proc gplot data=pieceout;
   plot (cost costhat)*lotsize/overlay;
```